# A STUDY OF ADOPTING BIG DATA TO CLOUD COMPUTING

ASMAA IBRAHIM
Technology Innovation and Entrepreneurship Center, Egypt
aelrehim@itida.gov.eg

MOHAMED EL-NAWAWY
Technology Innovation and Entrepreneurship Center, Egypt
melnawawy@itida.gov.eg



## ABSTRACT

Big data involve data of large volume and complex structure. The greatest challenge is to analyze these data in real time. One tool for achieving this is Apache Hadoop tool which could be used to provide a processing framework for OpenStack Cloud computing software tool. This paper presents a case study for fast provisioning of Hadoop clusters on OpenStack Development and Quality Assurance. In addition, Hadoop maximizes utilization of unused computational power for general purpose OpenStack IaaS cloud.Apache Hadoop and OpenStack represent two of the largest open source communities and both are relatively new to the data center. Hadoop can benefit from the operational agility provided by OpenStack and it serves as an excellent use case for OpenStack. A customized software, called sahara, is used for installing Hadoop in OpenStack. Thus, Hadoop APIs are integrated into OpenStack which has facilitated cluster management and considerably reduced time to manage cluster in terms of installation, troubleshooting, and operation. The core of Sahara, called the 'controller', serves as the glue between Hadoop and OpenStack. It manages the provisioning and orchestration of virtual machines by working with the underlying OpenStack projects like Nova, Quantum, Cinder and Glance.

## INTRODUCTION

With the creation of 2.5 quintillion bytes of data everyday where 90 percent of the data today were produced within the past two years (Nature Editorial,2008), the era of Big Data has arrived (Mervis et al.,2012), (Labrinidis et al.,2012), (IBM,2012). The current capability for data generation is unprecedented ever since the invention of the information technology in the early 19th century (Wu et al.,2014). The key challenge for Big Data applications is finding useful information from large volumes of data for future use (Rajaraman et al.,2011). Therefore, effective data analysis and prediction methods are needed in order to classify such volumes of data instantly.

Cloud computing is a new trend in information system and communication technology which promises economic benefits for industry, organizations and research community as migrating to cloud means applications will be located in data centers distributed across different continents rather than users' PCs (Safdari et al.,2014). Since cloud computing provides services through virtual resources and the internet, it allocates resources dynamically according to user needs and hence reduces costs of hardware as the application software is no longer installed on a local computer (Juan et al.,2014).

Currently, the focus is on a specific layer of Big Data ecosystem, e.g. OpenStack, (Wu et al., 2014). OpenStack is a free open source software project to build an IaaS (Infrastructure as a service) cloud deployment model (Robles et al., 2014). Since its first official release in 2010, over 200 companies have become involved in the project so far (Robles et al.,2014). In (Wu et al.,2014), the IaaS layer establishes a cluster to recognize vehicle license plates from images taken by the traffic cameras. While in (Tarannum et al.,2013), a Cloud computing framework for handling national ID database structure of Bangladesh is presented where an interactive graphical user interface for Bangladeshi People Search (BDPS) that uses a hybrid structure of cloud computing is proposed. Furthermore, (Yang et al.,2014), analyzes the specific application of cloud computing in e-commerce development and discusses how it processes data. In this paper, Hadoop is installed in OpenStack to create clusters for processing large volume data to create statistics of trainees in a training program.

Sahara, which is a customized software used for installing Hadoop in OpenStack, provides a scalable data processing stack and associated management interfaces by providing the ability to rapidly create and manage Apache Hadoop™ clusters and easily run workloads across them (Cohen, S., 2014). Thus, without having to deal with the details of cluster management, an All on OpenStack managed infrastructure is created.

With full cluster lifecycle management, provisioning, scaling and termination, Sahara allows the user to select different Hadoop versions, cluster topologies and node hardware details.

According to (Cohen, S., 2014), Sahara key features and use cases are:

- Fast and agile Hadoop cluster deployment
- An extensible framework for management and provisioning components
- Running Hadoop workloads in few clicks without expertise in Hadoop operations
- “Analytics as a Service” utilization of unused computational capacity for ad-hoc or bursty analytic workloads
- Sahara supports different types of jobs: MapReduce, Hive, Pig and Oozie workflows. The data could be taken from various sources: Swift, HDFS, NoSQL and SQL databases. It also supports various provisioning plugins.
- Represents an intersection of two of the largest open source movements
- OpenStack provides the foundation and hub of innovation for cleanly managing infrastructure resources. While Apache Hadoop™ serves as the core and innovation driver for storing and processing data.

**SETUP ENVIROMENT**

According to (openstack.org), the main setup environment for OpenStack involves a Three-node architecture:

- The OpenStack controller node contains all OpenStack Application Programming Interface services. These include integrating Hadoop as a component of OpenStack, all OpenStack schedulers, which are used to determine how to dispatch compute, network and volume requests, and Memcached service, which caches data and objects in memory to minimize the frequency with which an external database must be accessed.

- The OpenStack network node includes a set of application program interfaces (APIs), plug-ins and authentication/authorization control software that enables interoperability and orchestration of network devices and technologies within infrastructure-as-a-service (IaaS) environment for OpenStack.
- The OpenStack compute node is a server that can provide CPU, storage, network, and memory resources thus it greatly affects the utilization and performance of the cloud deployment model.

The number of compute nodes used was 7. Fig. 1 shows the Three-node architecture.

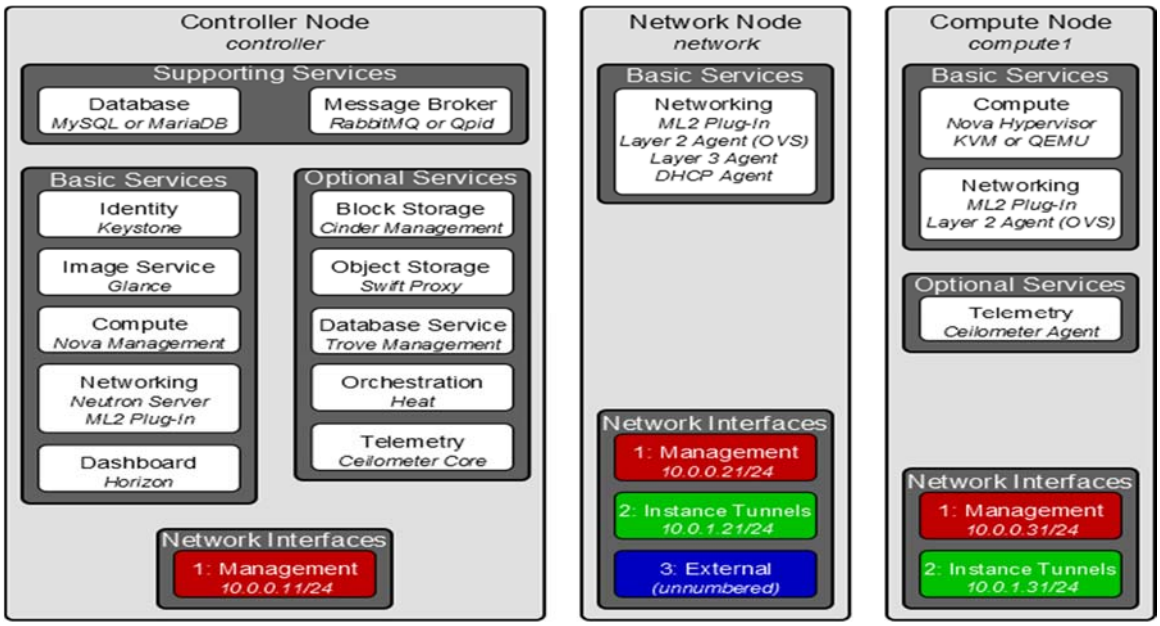


*Figure 1: Three-node architecture, Source: http://docs.openstack.org/juno/install-guide/install/yum/content/ch_overview.html#architecture_conceptual-architecture*

For Hadoop installation, we used the Data processing service for OpenStack, Sahara, which provides means to provision Hadoop clusters by specifying several parameters like Hadoop version, cluster topology, nodes hardware details and a few more. After user fills in all the parameters, the Data processing service deploys the cluster in a few minutes, in addition, sahara provides means to scale already provisioned clusters by adding/removing worker nodes on demand (Openstack.org).

According to (Openstack.org), the Sahara architecture consists of several components:

- The first is Auth component, which is responsible for client authentication & authorization.
- The second is Data Access Layer, DAL, which provides simplified access to data stored in the database.
- Provisioning Engine which is the component responsible for communication with Nova, Heat, Cinder and Glance which are components of the OpenStack controller node.
- Vendor Plugins which is a pluggable mechanism responsible for configuring and launching Hadoop on provisioned Virtual Machines.

- Elastic Data Processing (EDP) responsible for scheduling and managing Hadoop jobs on clusters provisioned by Sahara.
- Representational state transfer (REST) API which exposes Sahara functionality allowing increased performance in terms of components interaction, scalability to support large numbers of components, simplicity of interfaces, modifiability of components to meet changing needs, visibility of communication between components by service agents, portability of components by moving program code with the data, and reliability in terms of resistance to failure at the system level in the presence of failures within components (www.wikipedia.org), (Fielding,2000).
- Python Sahara Client enabling users to perform most of the existing operations like retrieving template lists, creating Clusters, etc.
- Last, Sahara pages which is the GUI for the Sahara.

Fig. 2 illustrates the Sahara architecture.

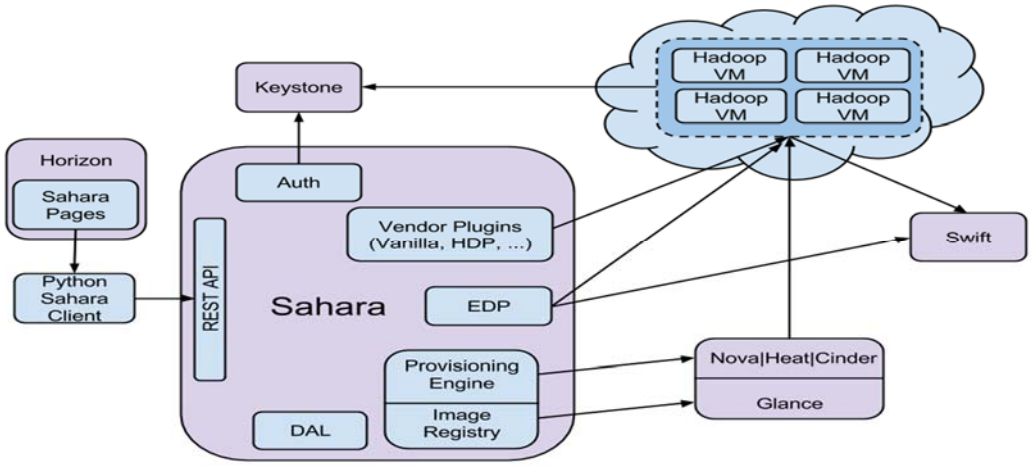


*Figure 2: Sahara architecture, Source: http://docs.openstack.org/developer/sahara/architecture.html*

As mentioned in (Openstack.org), once Hadoop has been installed, clusters can be created easily where a cluster deployed by Sahara consists of node groups that vary by their role, parameters and number of machines. Fig. 3 below illustrates an example of a Hadoop cluster consisting of 3 node groups each having a different role.

Based on the above setup environment, the environment we used included 1 Master group, with 1 JobTracker, and 1 Core Workers group, with 3 DataNodes, and without Workers group.

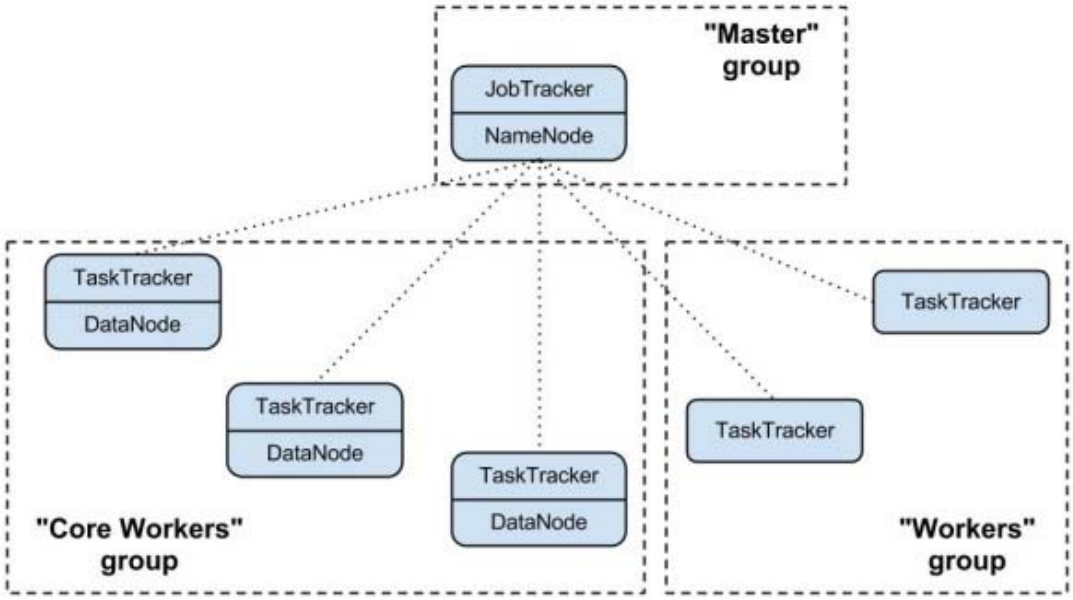


*Figure 3: Example of Hadoop cluster, Source: http://docs.openstack.org/developer/sahara/userdoc/overview.html*

**SAHARA: SYSTEM ADMINISTRATOR USE CASE**

As system administrators deal with the problems of installation, management, and monitoring of clusters, they need tools that are integrated into a central point of administration to monitor the entire IT infrastructure of the company (Kuznetzov, 2013).

As Hadoop provides enterprise-level tools, such as the Cloudera Management Console and Ambari for Apache/HortonWorks, another issue associated with a big organization is how to support different versions and distributions of Hadoop, where one team may prefer the Cloudera distribution, while another may prefer using a new feature that is currently available only in the latest version directly out of Apache (Kuznetzov, 2013).

Hadoop Cloud approach, matches the current tendency of moving applications to the cloud to enhance resources utilization(Kuznetzov, 2013) .

Through providing an easy way to set up multiple Hadoop environments, system administrators can choose OpenStack cloud as a tool for solving use-case-related issues(Kuznetzov, 2013) .

**APPLICATION**

Hadoop clusters provide utilization of unused computational power as well as Analytics as a Service. In the presented case, Hadoop clusters have been used to generate statistical output, using a development software tool named R, of a specific category of trainees in a training program, which has been conducted in several rounds over a period of 5 years. The data of the program is 4 Tera Bytes and the output was generated efficiently in few minutes.

## CONCLUSION

Hadoop clusters have been used with OpenStack for efficient utilization of high computational power to process a large amount of data where analytics for the processed data have been generated in few minutes. It has been shown that the process of Hadoop installation in OpenStack is simple using the Data processing service for OpenStack, Sahara.

## REFERENCES

Cohen, S., (2014) “Sahara: OpenStack Elastic Hadoop on Demand,”

Chapter 2 of Fielding's dissertation, (2000), "Network-based Application Architectures".

“IBM What Is Big Data: Bring Big Data to the Enterprise,” http:// www-01.ibm.com/software/data/bigdata/, IBM, 2012.

Juan, Q. J. and Jie, Y., (2014), "Laboratory Construction and management of university based on cloud computing," 3rd International Conference on Science and Social Research.

Kuznetzov, A., (2013) “Introducing Savanna, Hadoop as a Service for Private Cloud with OpenStack,”

Labrinidis, A.,(2012) and Jagadish, H., (2012), “Challenges and Opportunities with Big Data,” Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033.

Mervis, J., (2012), “U.S. Science Policy: Agencies Rally to Tackle Big Data,” Science, vol. 336, no. 6077, p. 22.

Nature Editorial, (2008), “Community Cleverness Required,” Nature, vol. 455, no. 7209, p. 1.

Rajaraman, A., and Ullman, J., (2011), Mining of Massive Data Sets. Cambridge Univ. Press.

Robles, G., Barahona, J., Cervigon, C., Capiluppi, A., Cortazar, D., (2014), "Estimating Development Effort in Free/Open Source software projects by mining software repositories: A case study of OpenStack," Proceedings of the 11th Working Conference on Mining Software Repositories.

Safdari, F. and Chang, V., (2014), "Review and analysis of cloud computing quality of experience," .

Tarannum, N., and Ahmed, N.,(2013), "EFFICIENT AND RELIABLE HYBRID CLOUD ARCHITECTURE FOR BIG DATABASE," International Journal on Cloud Computing: Services and Architecture (IJCCSA) ,Vol.3, No.6.

Wu, S., Chen, C., Chen,G., Chen, K.,Shou,L., Cao,H., Bai,H.,(2014), "YZStack: Provisioning Customizable Solution for Big Data," Proceedings of the VLDB Endowment, Vol. 7, No. 13.

Wu, X., Zhu, X., Wu, G., and Ding, W. (2014), "Data Mining with Big Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1

Yang,R., Lang,C.,(2014) "On Effects of Cloud Computing on Big Data Processing for E-commerce Development," 3rd International Conference on Science and Social Research (ICSSR 2014).

http://docs.openstack.org/juno/install-guide/install/yum/content/ch_overview.html#architecture_conceptual-architecture

http://docs.openstack.org/developer/sahara/architecture.html

http://docs.openstack.org/developer/sahara/userdoc/overview.html